

# Thermalization, freeze-out and noise: deciphering experimental quantum annealers

Jeffrey Marshall,<sup>1</sup> Eleanor G. Rieffel,<sup>2</sup> and Itay Hen<sup>1,3</sup>

<sup>1</sup>*Department of Physics and Astronomy, and Center for Quantum Information Science & Technology,  
University of Southern California, Los Angeles, California 90089, USA*

<sup>2</sup>*QuAIL, NASA Ames Research Center, Moffett Field, California 94035, USA*

<sup>3</sup>*Information Sciences Institute, University of Southern California, Marina del Rey, California 90292, USA*

By contrasting the performance of two quantum annealers operating at different temperatures, we address recent questions related to the role of temperature in these devices and their function as ‘Boltzmann samplers’. Using a method to reliably calculate the degeneracies of the energy levels of large-scale spin-glass instances, we are able to estimate the instance-dependent effective temperature from the output of annealing runs. Our results show that the output distributions of the annealers do not in general correspond to classical Boltzmann distributions. For the small fraction of the instances for which classical thermalization takes place, we find that the effective temperatures are significantly higher than the physical temperatures. Our results in this regime provide further evidence for the ‘freeze-out’ picture in which the output is sampled from equilibrium distributions determined at a point in time earlier in the quantum annealing process. We also find that the effective temperatures at different programming cycles fluctuate greatly, with the effect worsening with problem size. We discuss the implications of our results for the design of future quantum annealers to act as efficient Boltzmann samplers and for the programming of such annealers.

**Introduction.**— A handful of recent studies suggest that quantum annealers may be well suited to function as fast *thermal samplers* [1–4]. By taking advantage of their finite temperature nature [3–7], potentially they may sample from Boltzmann distributions of certain cost functions more efficiently than can be done classically. Such a capability opens up the exciting possibility of applications of quantum annealing to so-far-uncharted avenues of research, with immediate applications to domains such as deep learning networks and restricted Boltzmann machines [2, 3, 8].

The main mechanisms that determine the distributions from which output configurations are drawn are thus far unclear. Further insights into the role of temperature, and the capabilities of experimental quantum annealing optimizers to quickly thermalize, are challenging to gain due to the limited ability to probe the inner workings of these machines as well as the lack of control over most operating parameters that determine their dynamics, including temperature and annealing profiles [3, 4, 8].

To circumvent these difficulties, we have devised an experiment, directly comparing the performance of *two* commercially available quantum annealers operating at different temperatures (we shall refer to those as ‘hot’ and ‘cold’ henceforth). This key difference, together with a newly devised method to accurately calculate the degeneracies of large-scale spin-glass instances, offers us a unique opportunity to study the effects of temperature. Our results show that large-scale spin glasses do not in general equilibrate at classical Boltzmann distributions but are significantly affected by nonzero quantum fluctuations and noise. Our results provide corroboration to the proposed ‘freeze-out’ picture [1, 2, 9], which suggests that a quantum annealer is likely to experience a quasi-static evolution, returning a final population that

is close to a Boltzmann distribution of the final Hamiltonian but at an ‘effective (classical) temperature’ corresponding to a generally unknown point (or a narrow region) midway through the anneal where thermal, but generally not quantum, fluctuations become suppressed.

We also find that these effective temperatures fluctuate greatly at different programming cycles, with the effect worsening with problem size. We discuss factors potentially contributing to this adverse effect, including  $J$ -chaos in which control errors and noise mean that the problem run on the machine is different from the one programmed in. We discuss the implications of our results for the design of future quantum annealers to act as efficient Boltzmann samplers and for the programming of such annealers.

**Quantum annealing and quantum annealers.**— Standard transverse field quantum annealing works by evolving the system over rescaled time  $s = t/\mathcal{T} \in [0, 1]$  where  $t$  is time and  $\mathcal{T}$  is the overall runtime of the annealing process. The total Hamiltonian of the system is given by

$$H(s) = A(s)H_d + B(s)H_p, \quad (1)$$

where  $H_p = \sum_{\langle i,j \rangle} J_{ij} \sigma_i^z \sigma_j^z + \sum_i h_i \sigma_i^z$  is the programmable Ising spin-glass problem to be sampled defined by the parameters  $\{J_{ij}, h_i\}$ , and  $H_d = \sum_i \sigma_i^x$  is a transverse-field Hamiltonian which provides the quantum fluctuations. We identify two dimensionless scales associated with the annealing, namely, the one associated with quantum fluctuations  $Q(s) = A(s)/B(s)$  and the scale associated with thermal fluctuations  $k_B T/B(s)$ , both of which are shown in Fig. 1 for both the ‘hot’ and ‘cold’ processors.

Current quantum annealers suffer from intrinsic control errors (ICE) [6, 10] such as imperfect digital-to-analog conversion when programming the coupling pa-

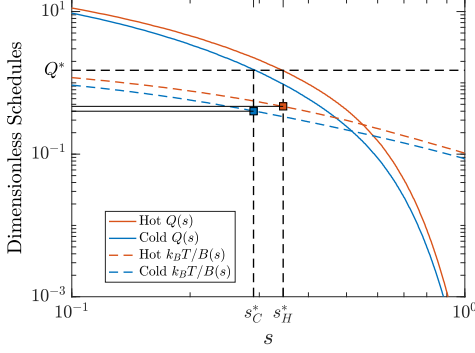


FIG. 1. **Dimensionless annealing schedules and temperatures of the hot and cold contrasted DW2 processors.** We plot the quantum fluctuations  $Q(s)$ , and the thermal fluctuations  $k_B T/B(s)$ , as a function of rescaled annealing time,  $s$ , for each machine. An example of freeze-out is shown; the distribution approximately halts at some fixed instance-dependent value of  $Q^* := Q(s^*)$  (dashed black line) which corresponds to a unique freeze-out point  $s^*$  (denoted  $s_C^*, s_H^*$ ) and dimensionless effective temperature  $k_B T/B(s^*)$  (squares), for each machine. As indicated by the solid black lines, freeze-out may correspond to differing effective temperatures.

rameters  $\{J_{ij}\}$  and magnetic fields  $\{h_i\}$  onto the machine, and  $1/f$ -noise whose effect is changing the parameters even within a single anneal [11, 12]. For both contrasted quantum annealers, these random errors may be approximated as normally distributed according to  $\sim \mathcal{N}(0, 0.05J)$  [resp.  $\sim \mathcal{N}(0, 0.03h)$ ] where  $J$  (resp.  $h$ ) is the maximal value over all the programmed  $J_{ij}$  (resp.  $h_i$ ). Some problems have resilience to such errors [5, 13], whereas others are susceptible to a phenomenon referred to as  $J$ -chaos, in which output ‘solutions’ correspond to the wrong, or malformed, problem, generally reducing the success probability [6, 13–17].

**Freeze-out conjecture.**— One may naturally expect quantum annealers to return configurations sampled from a Boltzmann distribution, in which each configuration  $c$  has weight proportional to  $e^{-\beta_{\text{ideal}}^{\text{eff}} E_c}$ , where  $E_c$  is the configuration’s classical cost and  $\beta_{\text{ideal}}^{\text{eff}} \equiv B(1)/k_B T$  is an effective dimensionless inverse temperature, with  $T$  being the operating temperature of the machine. The freeze-out conjecture provides a mechanism to explain for high observed effective temperatures [1, 2, 9]; the temperature of the Boltzmann distribution that best fits the distribution of results returned by the quantum annealer. The freeze-out conjecture suggests that observed higher effective temperatures indicate a freezing of the evolution at an unknown, instance dependent ‘freeze-out’ point  $s^*$  during the anneal, at which point the thermal fluctuations, whose strength is coupled to the quantum fluctuations  $Q(s^*)$  driving the system, become negligibly small [18]. As illustrated in Fig. 1, the freeze-out point

is conjectured to happen at a temperature-independent (but instance-dependent) value  $s^*$  [1]. Only when  $Q(s^*)$  at the freeze-out is small is the final distribution expected to be a classical Boltzmann distribution for  $H_p$  with (dimensionless) effective temperature  $\beta^{\text{eff}} = B(s^*)/k_B T$ ; otherwise, the resultant distribution generally will not correspond to an equilibration at any given point, but instead result from different parts of the system equilibrating at different temperatures and times [1].

**Experimental setup.**— We made use of two 512-qubit D-Wave Two (DW2) quantum annealers [19], where the mean temperature of the ‘hot’ (‘cold’) machine was about 16.0 mK (13.2 mK) [20]. To test the processors, we designed 1300 random spin-glass instances of the planted-solution type [21] for each of seven different problem sizes corresponding to  $L \times L$  grids of 8-qubit cells of the hardware DW2 Chimera graph with  $L = 2 \dots 8$  (see Fig. 1 in the SI). This class of instances is particularly suitable for our purposes for two main reasons: i) the ground state energies of the generated problems are known in advance, and ii) the *exact* degeneracies of the ground and first excited states are computable [4]. We generate instances on the intersection of the two hardware graphs (501 qubits) in order to avoid biases associated with malfunctioning qubits on either machine. We ran each instance with 440,000 anneals over twenty two programming cycles each consisting of up to  $N_{\text{anneals}} = 20,000$  anneals per cycle and with anneal times in the range  $[20, 40]\mu\text{s}$  [22].

**Methods: degeneracy counting.**— To evaluate  $\beta^{\text{eff}}$ , we must estimate the degeneracies of the energy levels of the problem instances with high accuracy and confidence. To do so, we employ two independent, complementary, techniques. The first is the well-known Wang-Landau (WL) entropic sampler [23] (see Fig. 4 in the SI), which statistically estimates the degeneracy of the energy levels. Since the WL algorithm is prone to statistical errors as well as false convergences, we employ in parallel a newly devised algorithm which takes advantage of the fact that planted-solution instances can be written as a sum of local terms [4]. The algorithm allows us to compute the degeneracies of the ground and first excited states *exactly*. We then check for how close the WL estimates come to these exact values. If the WL estimate is outside  $\pm 5\%$  of the exact value for either the ground or first excited state, we discard this instance as we know it has not converged properly. The combination of the two algorithms allows for the faithful estimation of the degeneracies for some 2200 instances in total, of differing problem sizes (see SI).

Armed with the degeneracies, we estimate  $\beta^{\text{eff}}$  by minimizing

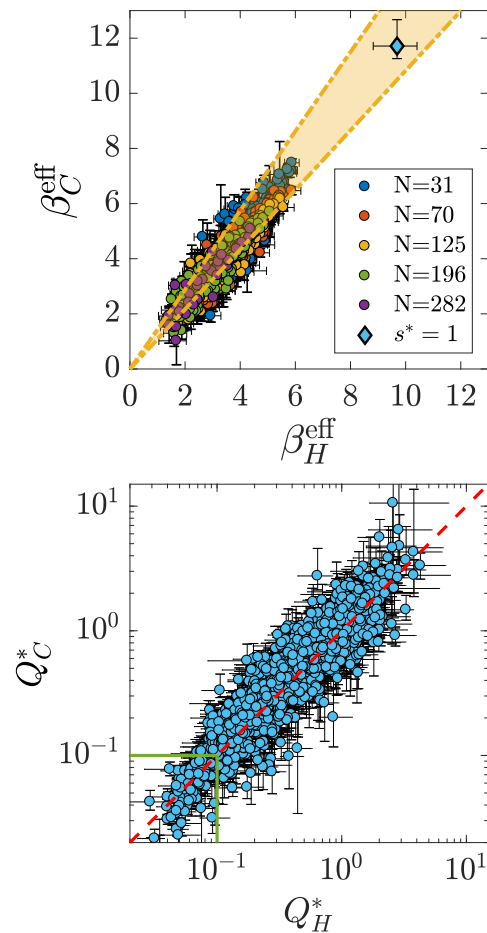
$$\left| P_0 - \left( \sum_{k=0} g_k e^{-\beta^{\text{eff}} (E_k - E_0)} \right)^{-1} \right|, \quad (2)$$

where  $P_0$  is the ground state success probability obtained from the DW2, and  $\{g_k, E_k\}$  are the degeneracy and energy of the  $k$ -th level, respectively. Agreement between the two methods, and considering instances which DW2 could successfully solve, enabled us to accurately estimate  $\beta^{\text{eff}}$  for problem sizes of up to 282 qubits [24].

**Results: thermalization and freeze-out.**— In Fig. 2(top), we plot the median inverse temperature  $\beta^{\text{eff}}$  for each instance and machine (error bars represent the maximum and minimum value of  $\beta^{\text{eff}}$  found over all programming cycles). Evident is the overall strong linear correlation between the (inverse) temperatures of the two machines, indicating correlated performances with Pearson coefficient 0.94. Most of the instances fall within, or near, the ‘thermal’ range predicted by the ratio of physical temperatures of the machines [see yellow band in Fig. 2(top)], illustrating the key functional role of temperature in the success probability of these problems. Note however, that if indeed the instances were thermalizing at the end of the anneal, we would expect to observe  $\beta_{\text{ideal}}^{\text{eff}}$  of 9.7 and 11.7 for the hot and cold machines, respectively (also shown in the figure). Instead, the values we observe are well below this mark:  $\beta^{\text{eff}} \in [2, 7]$ . Thus, we are finding effective temperatures up to six times higher than would be expected from a simple thermalization picture. Moreover, the ratio of  $\beta^{\text{eff}}$  for the two machines,  $R_\beta = 1.14$ , is well below the ratio of the physical temperatures,  $R_\beta^{\text{ideal}} = 11.7/9.7 \approx 1.21$  indicating an effective average temperature ratio of about 94% of the ‘thermal’ ratio of  $s = 1$ . We now examine the extent to which the freeze-out picture can explain these discrepancies.

**Analysis: freeze-out.**— Having two machines at different operating temperatures enables us to test the freeze-out conjecture. While the freeze-out point for each instance is unknown, its temperature independence means the estimates for the freeze out point should be the same whether based on data from the cold machine or the hot machine. Using the estimated  $\beta^{\text{eff}} = B(s^*)/k_B T$ , from which we can obtain the freeze-out point  $s^*$  given the known operating temperatures and annealing schedules, we directly calculate  $Q(s^*)$ . This in turn allows us to check whether indeed the freeze-out point is the same for the identical instances [that is, whether  $Q(s^*)$  is the same at the freeze-out point]. We plot  $Q(s^*)$  for each instance in Fig. 2(bottom). The smaller the value of  $Q(s)$ , the more likely it is to correspond to classical final Boltzmann distributions (and therefore also meaningful  $\beta^{\text{eff}}$  calculations). When we restrict to instances with small  $Q$  (we take  $Q < 10^{-1}$ ), we find excellent correspondence, with an average ratio of  $R_Q^{\text{small}} = 1.01 \pm 0.06$  (95% confidence interval), in agreement with the freeze-out hypothesis.

As is clear from the figure, only a small fraction of the instances correspond to a negligible  $Q(s^*)$ . Overall we observe that the ratio is substantially higher,  $R_Q \approx 1.20$ ,



**FIG. 2. Top: Effective inverse temperatures.** Comparison of effective inverse temperatures on the set of instances for both the hot processor, denoted by  $\beta_H^{\text{eff}}$ , and the cold processor, denoted by  $\beta_C^{\text{eff}}$ . Error bars represent the highest and lowest values found over all programming cycles. The yellow band represents the range of physical temperature fluctuations between the devices (see Fig. 3 of SI). The blue diamond represents  $\beta^{\text{eff}}$  were (classically) thermalization to take place at the end of the anneal. Differently colored points denote different problem sizes (see legend). **Bottom: Quantum fluctuations at freeze-out.** Scatter plot of the extracted  $Q^* := Q(s^*)$  strengths (blue) for each instance. Ideally, both machines should yield the same value for each instance (red  $y = x$  line). Median ratio is  $R_Q^{\text{small}} = 1.01$  for small  $Q(s^*)$  values (within the green square). The outputs are strongly correlated with a Pearson coefficient of 0.92. Error bars represent the range of  $Q^*$  values over all programming cycles.

and deviates further and further from the ‘ideal’ value of 1 as  $Q$  increases. It is worth comparing small  $Q(s^*)$  instances with the rest of the instances. The small  $Q(s^*)$  problems are typically easier to solve and are proportionately smaller in problem size than the bulk of the instances we tested. Indeed, for most of the instances, the calculated  $Q(s^*)$  values correspond to distributions that are far from classical. Interestingly, the freeze-out

picture is also consistent with the lower-than-ideal effective inverse-temperature ratio  $R_\beta = 1.14$  (and in turn a higher  $R_Q \approx 1.20$ ). The existence of significant quantum fluctuations in both machines leads to an overestimation of thermal fluctuations, i.e., to higher effective temperatures. Since the quantum fluctuations in the freeze-out point are comparable on both machines, the resultant ratio is expected to be lower, as we indeed observe [25].

**Analysis: high variability in inverse temperature estimates.**— The magnitude of the error bars on the effective inverse temperatures per instance shown in Fig. 2(top) reflect the large fluctuations in success probabilities between programming cycles. We discuss various factors that contribute to that variance.

It is known that the location of the freeze-out point (and hence the success probability) has a weak logarithmic dependence on the annealing time [1, 6], with longer anneal times having later freeze-out points because there is more time for fluctuations to take place. We indeed find such an effect (see Figs. 7 and 8 of the SI), though our results show that this typically accounts for less than a 1% variability between different anneal times and therefore does not explain the drastic spread we observe. If the variation were due to purely statistical variations from cycle to cycle, one would expect statistical fluctuations in success probability  $P_0$  on the order of  $\delta P_0 = \sqrt{P_0(1 - P_0)/N_{\text{anneals}}}$ . Fig. 3(left) shows  $R_{\Delta/\delta} = \Delta P_0 / \delta P_0$ , the ratio of typical magnitude of actual fluctuations in success probabilities  $\Delta P_0$  to the expected magnitude of purely statistical fluctuations  $\delta P_0$ . We find that only around 20% of the instances exhibit fluctuations of success probability  $R_{\Delta/\delta}$  below 1. For most instances, typical fluctuations are about an order of magnitude greater than statistical fluctuations, with some fluctuations being considerably greater. We attribute these large ratios, to  $J$ -chaos [6] from ICE and other noise, which affect the local fields and coupling parameters within and between cycles. Noise unrelated to programming parameters may also play a role.

Figure 3(right) shows, as a function of problem size, the average variation in  $\beta^{\text{eff}}$ , as measured by the ratio of the 95th to 5th percentile values found over all programming cycles. The trend is clear; the larger the problem size, the greater the size of the fluctuations. It is critical to understand why these fluctuations scale with problem size, and their root cause, so as to devise strategies to keep these errors from becoming unmanageable as chip sizes increase. For a fixed problem size, we do not observe a clear correlation between success probability and the variance in the  $\beta^{\text{eff}}$  estimates (Fig. 9 of the SI), providing evidence that the fluctuations we observe in Fig. 3(right) are indeed due to differences in problem size and not problem difficulty (though of course the two are related) [26]. This trend is expected as larger problems, with more couplings, have more potential to be adversely affected by control errors, and other sources of

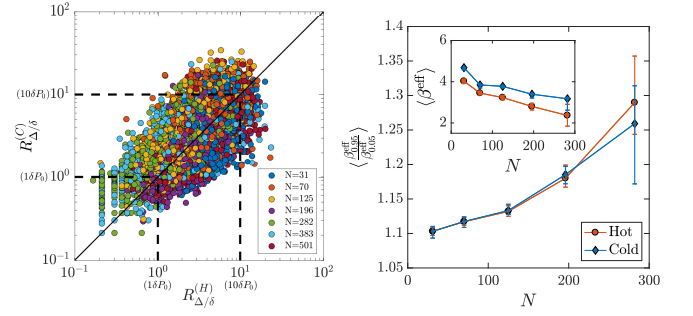


FIG. 3. **Left: Ratio of magnitude of actual fluctuations in success probabilities to magnitude of statistical fluctuations,  $R_{\Delta/\delta}$ , for the various instances on the hot and cold processors.** For most instances, the fluctuations in success probabilities over programming cycles is far greater (by an order of magnitude) than what one might expect from fluctuations of purely statistical nature. **Right: Typical spread of effective inverse temperature as a function of problem size.** Our measure for spread is the 95th to 5th percentile mean ratio of  $\beta^{\text{eff}}$  averaged over instances of each problem size. We take the ratio to overcome any bias resulting from the fact the cold chip records higher values of  $\beta^{\text{eff}}$  (see inset). Remarkably, we find both devices follow a nearly identical trend: fluctuations increase and  $\beta^{\text{eff}}$  decreases with problem size. Inset: Median  $\beta^{\text{eff}}$  for each problem size.

noise [27].

**Conclusions.**— By conducting parallel experiments on two quantum annealers, each operating at a different temperature, we studied key mechanisms determining their output distributions. Our results provide corroboration to the freeze-out conjecture [1, 2, 9] for some problems. To test this conjecture, we compared the performance of the two machines on certain Ising problems, making use of a recent method to accurately estimate the degeneracies of such problems. For instances in which the freeze-out point corresponds to negligible quantum fluctuations, the effective temperature at which (classical) thermalization takes place is both instance dependent and independent of the operating temperature of the device. These results suggest a strong need for further research into the extent to which D-Wave machines, and quantum annealers more generally, can function as efficient Boltzmann samplers.

Our results also show, however, that thermalization to the classical Boltzmann distributions takes place for only a small fraction of instances. Instead, most instances do not equilibrate uniformly at a specific point. Moreover, we have found that the effective temperatures at different programming cycles can wildly fluctuate. Our data indicates that this effect worsens with larger problem size and thus calls into question the use of quantum annealers as classical Boltzmann samplers. Our observations need to be taken into account as researchers work to design and build bigger and better quantum annealers.



Promising directions include reducing the various sources of noise that contribute to intrinsic control errors (ICE) in quantum annealing hardware, and exploring the potential for alternate annealing schedules and non-standard drivers to enable more instances to equilibrate at a unique point late enough in the anneal that the quantum fluctuations are negligible. For machine learning, and other applications, it is not clear how accurately one needs to sample from a Boltzmann distribution, or even that Boltzmann distributions are optimal for this purpose. Another tantalizing direction to pursue is possible use of the distributions that have a large quantum component [2], particularly given that certain distributions generated by quantum Hamiltonians are believed to have no efficient classical sampling mechanism [28, 29]. A deeper understanding of these processes will have profound implications for the design of future annealers and the prospects of utilizing quantum annealers as efficient Boltzmann samplers for machine learning and beyond.

**Acknowledgments.**—We thank Tameem Albash, Mohammad Amin, Salvatore Mandrà and Walter Vinci for useful discussions. Part of the computing resources were provided by the USC Center for High Performance Computing and Communications and the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. ER would like to acknowledge support from the NASA Advanced Exploration Systems program and NASA Ames Research Center. Her contributions to this work were also supported in part by the AFRL Information Directorate under grant F4HBKC4162G001 and the Office of the Director of National Intelligence (ODNI). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, AFRL, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purpose notwithstanding any copyright annotation thereon.

---

[1] M. H. Amin, *Phys. Rev. A* **92**, 052323 (2015).  
 [2] M. H. Amin, E. Andriyash, J. Rolfe, B. Kulchytskyi, R. Melko, ArXiv e-prints (2016), [arXiv:1601.02036 \[quant-ph\]](#).  
 [3] M. Benedetti, J. Realpe-Gómez, R. Biswas, and A. Perdomo-Ortiz, *Phys. Rev. A* **94**, 022308 (2016).  
 [4] B. H. Zhang, G. Wagenbreth, V. Martin-Mayor, and I. Hen, ArXiv e-prints (2017), [arXiv:1701.01524 \[quant-ph\]](#).  
 [5] H. G. Katzgraber, F. Hamze, Z. Zhu, A. J. Ochoa, and H. Muñoz-Bauza, *Phys. Rev. X* **5**, 031026 (2015).  
 [6] V. Martin-Mayor and I. Hen, *Scientific Reports* **5**, 15324 (2015).

[7] J. Marshall, V. Martin-Mayor, and I. Hen, *Phys. Rev. A* **94**, 012320 (2016).  
 [8] S. H. Adachi and M. P. Henderson, ArXiv e-prints (2015), [arXiv:1510.06356 \[quant-ph\]](#).  
 [9] M. W. Johnson *et al.*, *Nature* **473**, 194 (2011).  
 [10] A. D. King and C. C. McGeoch, ArXiv e-prints (2014), [arXiv:1410.2628 \[quant-ph\]](#).  
 [11] Z. Zhu, A. J. Ochoa, S. Schnabel, F. Hamze, and H. G. Katzgraber, *Phys. Rev. A* **93**, 012317 (2016).  
 [12] S. Boixo, V. N. Smelyanskiy, A. Shabani, S. V. Isakov, M. Dykman, V. S. Denchev, M. H. Amin, A. Y. Smirnov, M. Mohseni, and H. Neven, *Nature Communications* **7**, 10327 EP (2016).  
 [13] D. Venturelli, S. Mandrà, S. Knysh, B. O’Gorman, R. Biswas, and V. Smelyanskiy, *Phys. Rev. X* **5**, 031040 (2015).  
 [14] M. Nifle and H. J. Hilhorst, *Phys. Rev. Lett.* **68**, 2992 (1992).  
 [15] M. Ney-Nifle, *Phys. Rev. B* **57**, 492 (1998).  
 [16] F. Krzakala and J. P. Bouchaud, *Europhys. Lett.* **72**, 472 (2005).  
 [17] H. G. Katzgraber and F. Krzakala, *Phys. Rev. Lett.* **98**, 017201 (2007).  
 [18] The term ‘effective temperature’ is somewhat of a misuse as it may imply thermalization of the system whereas in fact it may not be the case.  
 [19] One machine is owned by Lockheed-Martin, housed at University of Southern California’s Information Sciences Institute and the other, purchased by a NASA-USRA-Google collaboration located inside the NASA Ames Research Center.  
 [20] Further details on the two D-Wave processors are provided in the Supplemental Information (SI).  
 [21] I. Hen, J. Job, T. Albash, T. F. Rønnow, M. Troyer, and D. A. Lidar, *Phys. Rev. A* **92**, 042325 (2015).  
 [22] See SI for more information.  
 [23] F. Wang and D. P. Landau, *Phys. Rev. E* **64**, 056101 (2001).  
 [24] We were not able to obtain accurate estimates for  $\beta^{\text{eff}}$  for the largest problems,  $L = 7, 8$ . We discuss this in more detail in the SI.  
 [25] We also provide in the SI a density plot (‘heat map’), and plots for separate problem sizes for Fig. 2 (see Figs. 10-12 of the SI).  
 [26] We also discount any discrepancy due to the logarithmic dependence on anneal time (inset of Fig. 8 in the SI), as it is not correlated with problem size.  
 [27] The increase in fluctuations with problem size we observe in Fig. 3(right) is most likely an underestimate of the full effect. Since our criterion for discarding instances is convergence of the WL algorithm, it is very likely that those instances that do not appear in the figure exhibit fluctuations of larger magnitudes, as there is a known strong positive correlation between WL convergence, i.e., the classical hardness of an instance, and  $J$ -chaos (see, e.g., Ref. [6]).  
 [28] C. Semay and L. Ducobu, *European Journal of Physics* **37**, 045403 (2016).  
 [29] R. W. Robinett, *Am. J. Phys.* **63**, 823 (1995).  
 [30] P. I. Bunyk, E. M. Hoskinson, M. W. Johnson, E. Tolkacheva, F. Altomare, A. Berkley, R. Harris, J. P. Hilton, T. Lanting, A. Przybysz, and J. Whittaker, *Applied Superconductivity, IEEE Transactions on, Applied Superconductivity, IEEE Transactions on* **24**, 1 (Aug. 2014).

- [31] V. Choi, [Quant. Inf. Proc.](#) **10**, 343 (2011).
- [32] W. Barthel, A. K. Hartmann, M. Leone, F. Ricci-Tersenghi, M. Weigt, and R. Zecchina, [Phys. Rev. Lett.](#) **88**, 188701 (2002).
- [33] F. Krzakala and L. Zdeborová, [Phys. Rev. Lett.](#) **102**, 238701 (2009).
- [34] S. Boixo, T. F. Ronnow, S. V. Isakov, Z. Wang, D. Wecker, D. A. Lidar, J. M. Martinis, and M. Troyer, [Nat. Phys.](#) **10**, 218 (2014).
- [35] T. F. Rønnow, Z. Wang, J. Job, S. Boixo, S. V. Isakov, D. Wecker, J. M. Martinis, D. A. Lidar, and M. Troyer, [Science](#) **345**, 420 (2014).
- [36] M. Mezard, G. Parisi and M.A. Virasoro, *Spin Glass Theory and Beyond*, World Scientific Lecture Notes in Physics (World Scientific, Singapore, 1987).
- [37] In fact, we find that the WL algorithm does indeed find it very hard to converge properly on instance sizes greater than 282 bits.

## SUPPLEMENTAL INFORMATION

### The Google-NASA-USRA ('cold') and Lockheed-Martin-USC ('hot') D-Wave Two processors

The putative quantum annealer used in our work is the D-Wave Two (DW2) device [30]. This device is designed to solve optimization problems by evolving a known initial configuration — the ground state of a transverse field  $H_d = \sum_i \sigma_i^x$ , where  $\sigma_i^x$  is the Pauli spin-1/2 matrix acting on spin  $i$  — towards the ground state of a classical Ising-model Hamiltonian which serves as a cost function that encodes the problem that is to be solved:

$$H_p = \sum_{\langle i,j \rangle} J_{ij} \sigma_i^z \sigma_j^z + \sum_i h_i \sigma_i^z. \quad (1)$$

The variables  $\{\sigma_i^z\}$  denote either classical Ising-spin variables that take values  $\pm 1$  or Pauli spin-1/2 matrices, the  $\{J_{ij}\}$  are programmable coupling parameters, and the  $\{h_i\}$  are programmable local longitudinal fields. The  $N$  spin variables are realized as superconducting flux qubits and occupy the vertices of the D-Wave 'Chimera' hardware graph [30, 31]. Here,  $\langle i,j \rangle$  denotes summation over the edges of the graph. The union of the two D-Wave Chimera graphs is given in Fig. 1 – this is the graph all of our problem instances were defined on.

These machines evolve the full Hamiltonian via

$$H(s) = A(s)H_d + B(s)H_p. \quad (2)$$

The way in which the strength of the initial ( $H_d$ ) and final ( $H_p$ ) Hamiltonians evolve is given by the parameters  $A(s)$  and  $B(s)$ , where  $s = t/T \in [0, 1]$  is the annealing time. Here,  $T$  is the total annealing time, ranging between  $20\mu\text{s}$  and  $20\text{ms}$  on these devices. The annealing schedule is given in Fig. 2 for each machine.

In Fig. 3 we show the temperature log of the D-Wave chips during the time which we collected our data.

### Generation of instances

For the generation of instances, we have chosen in this work to study problems constructed around 'planted solutions' – an idea borrowed from constraint satisfaction (SAT) problems [32, 33]. In these problems, the planted solution represents a ground state configuration of Eq. (1) that minimizes the energy and is known in advance. This knowledge circumvents the need to verify the ground state energy using exact (provable) solvers, which rapidly become too expensive computationally as the number of variables grows, and which were employed in earlier benchmarking studies [34, 35]. Moreover, these problems are known to possess different degrees of 'tunable hardness', achieved by adjusting the amount of frustration (see Ref. [36]) which we will use. Last, studying

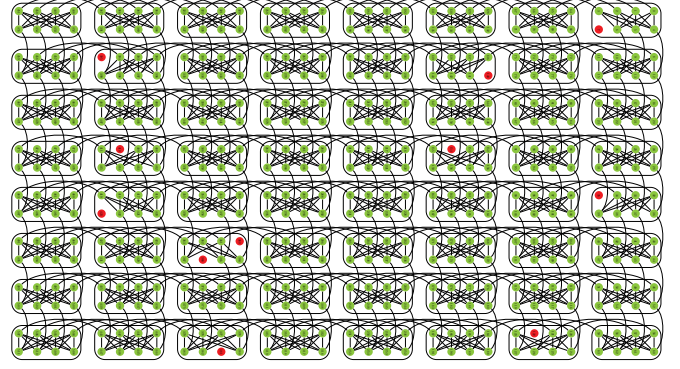


FIG. 1. **Intersected Chimera.** Intersection of the two D-Wave Chimera graphs, with 501 operating qubits. Red disks denote non-operational qubits on one (or both) of the two machines.

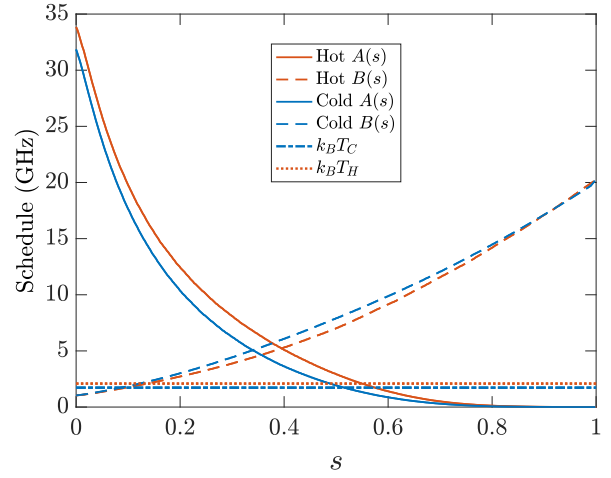


FIG. 2. **Annealing schedules of the USC (hot) and NASA (cold) DW2 processors.** Annealing schedule [see Eq. (2)] in GHz as a function of dimensionless annealing time  $s = t/T$ . We also plot the temperatures ( $\hbar = 1$ ) of the devices (see legend).

this type of problems will allow us to devise an algorithm to find all minimizing configurations of the generated instances. The interested reader will find a more detailed discussion of planted Ising problems in Ref. [21].

### Wang-Landau entropic sampler

As explained in the main text, we employed a Wang-Landau entropic sampler to estimate the degeneracy of the energy levels for our generated planted-solution instances. This algorithm performs essentially a random walk over the energy landscape, where updates at each step in the algorithm are such that an approximately flat

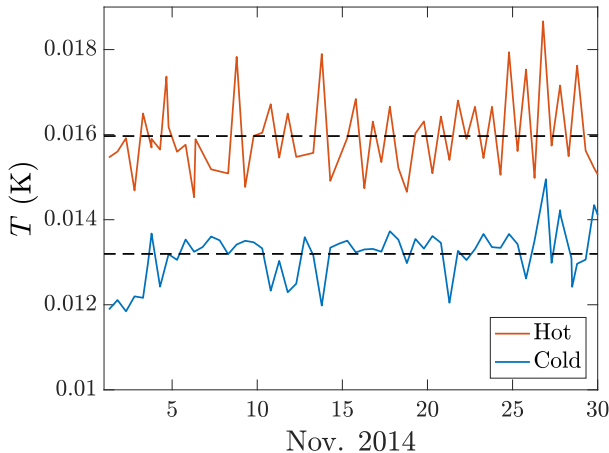


FIG. 3. **Temperature logs.** Temperature log for the USC (hot) and NASA (cold) machines during the period of time we performed our experiments. The dashed black lines represent the mean of the data sets. The mean temperature of the ‘hot’ USC machine was about  $T_H = 16.0$  mK, and the mean temperature of the ‘cold’ NASA machine was about  $T_C = 13.2$  mK, with ratio  $T_H/T_C \approx 1.21$ . Note, the temperature data is sparse, sampled only twice per 24 hours.

histogram of visited energies is produced. We follow the same methodology as originally described in [23]. Our histogram was considered ‘flat’ when the lowest sampled energy level has been visited at least 80% of the mean of the entire histogram.

We performed 20 independent Wang-Landau runs, each up to  $10^9$  steps for each of our instances. We then averaged over these 20 runs which provided our estimate of degeneracies for each instance. We then discarded any instances for which the ground or first excited state degeneracies did not match that for the exact solution counter (up to 5% error). This meant we had accurate degeneracy data for problems up to 282 qubits in size [37]. See Fig. 4.

The total number of instances for which  $\beta^{\text{eff}}$  was successfully estimated was, for each problem size  $L = 2 \dots 8$  is [664,745,449,266,38,0,0]. The difficulty in obtaining an accurate measurement, especially for the larger problems, was due mainly to i) the D-Wave machine not being able to solve many of the ‘hard’ problems and ii) there were too many degenerate states for the exact counter to enumerate (exceeded our chosen cut-off value of  $10^7$  ground states, which become prohibitively expensive to compute), or iii) Wang-Landau estimate deviated too far from exact counter results (generally from under-sampling the low energy states).

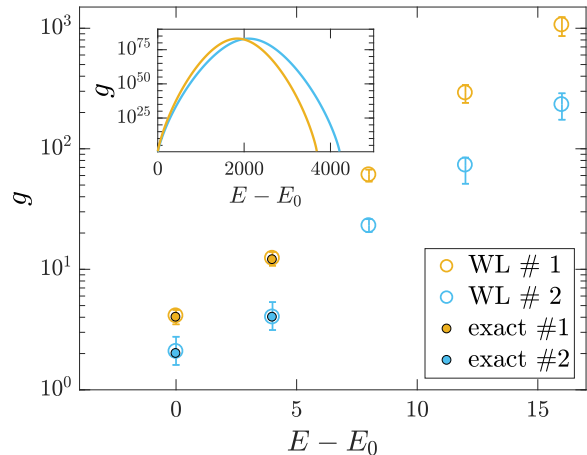


FIG. 4. **Degeneracy counting.** Main figure: The degeneracies of the first five energy levels of two of the 282-qubit problem instances as found by the Wang-Landau algorithm (error bars represent 95% confidence interval). The WL degeneracies of the first two levels lie on top of the computable exact values (solid circles). Inset: Degeneracies of all levels as a function of energy, for the same instances, as obtained by the WL algorithm.

### Estimation of $\beta^{\text{eff}}$

A central part of our analysis is the estimation of  $\beta^{\text{eff}}$ . As explained in the main text, we do this by minimizing

$$\left| P_0 - \left( \sum_{k=0} \frac{g_k}{g_0} e^{-\beta^{\text{eff}}(E_k - E_0)} \right)^{-1} \right|. \quad (3)$$

One important point to realize is that the D-Wave rescales all coupling values such that the encoded values,  $\tilde{J} \in [-1, 1]$ . The planted-solution instances have energies  $E_i = E_0 + 4n_i$  where  $i = 0, 1, \dots$  and  $n_i \in \mathbb{N}$  (ordered such that  $E_i > E_j \implies i > j$ ). The D-Wave rescaling means that the  $E_k - E_0$  appearing in Eq. (3) is not of the form  $4n_k$ , but actually  $E_k - E_0 = 4n_k/J_{\text{max}}$ , where  $J_{\text{MAX}} := \max_{(i,j) \in E} |J_{ij}|$  (i.e.,  $H_p \rightarrow H_p/J_{\text{max}}$ ). Note,  $E_i$  is dimensionless [the energy unit is carried by  $B(s)$ ].

### Data collection

We generated 13 groups of 100 instances, for each of 7 different qubit numbers  $L = 2 \dots 8$ , see Fig. 1 (i.e., 9100 total instances). These 13 groups differed in the number of clauses (or loops) contained in each instance. We ran each instance sequentially, with anneal times in range  $[20-40] \mu\text{s}$  (in steps of  $2 \mu\text{s}$ ). We then repeated this process, so that each instance was run over 22 programming cycles on each machine. Note, we did not gauge



average the instances. All of these instances were run during November of 2014.

In Fig. 5 we show the histogram of the success probabilities for the two machines, for all of the instances. We see the cold (NASA) machine clearly outperforms the hot (USC) machine—we expect, due to the colder operating temperature. In Fig. 6 we compare two different programming cycles (from different days) on the same machine, showing consistency over different runs.

We also study the effect of varying anneal time on success probability in Figs. 7 and 8. We see that there is only a very weak (logarithmic) dependence on anneal time, in accordance with [1, 6], and moreover, it is seemingly not correlated with problem size.

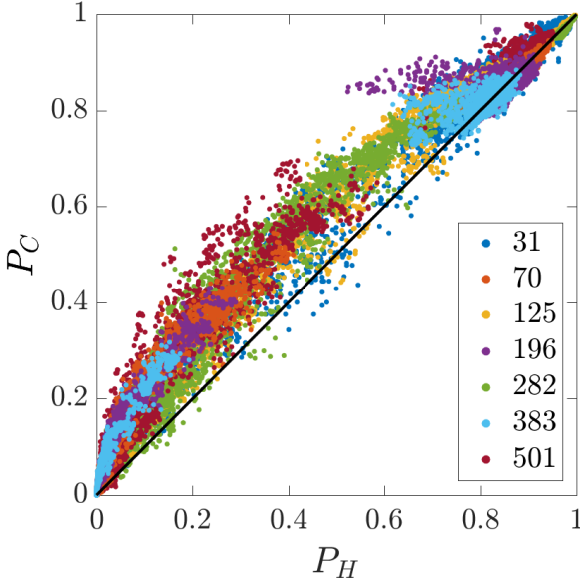


FIG. 5. **Success probability.** Probability of success (how often the ground state energy is correctly identified) of all instances and programming cycles on the two machines ('hot' USC machine  $P_H$ , and the 'cold' NASA machine  $P_C$ ). Each point is a randomly chosen programming cycle (for the same instance on each machine). Number of qubits given by legend.

#### Extra figures

In Fig. 9 we plot (relating to Fig. 3 of the main text), the typical spread in the effective inverse temperatures for all instances for problem size 70, against success probability (this is problem the size for which we have the most number of instances with accurate degeneracy data from Wang-Landau). We in fact see no clear correlation in the data, suggesting the trend we observe in Fig. 3 of the main text is indeed due to problem size, and not problem difficulty, per se.

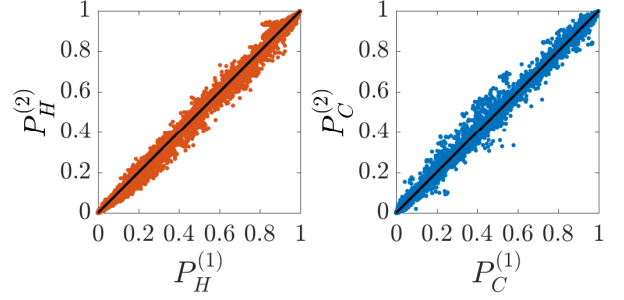


FIG. 6. **Machine correlation.** We compare the results of two programming cycles for each instance on each machine. We see the data aligns nicely along  $y = x$ , albeit with sizable fluctuations (as one would expect). Compare this with Fig. 5, where data clearly deviates from  $y = x$ . The 'hot' USC machine is on the left, and the 'cold' NASA machine on the right.

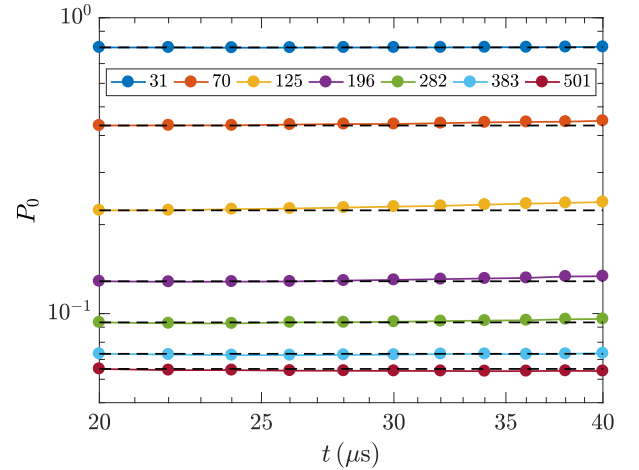


FIG. 7. **D-Wave success with anneal time.** Average probability of success,  $P_0$  (for the hotter USC machine), against anneal time (log scale), for different problem sizes (see legend). Each point averaged over two programming cycles, of  $N_{\text{anneals}} = 20,000$  anneals each. The black dash lines correspond to the value of  $P_0$  for  $t = 20 \mu\text{s}$ . One can see in general a slight increase in success probability with  $t$ .

Relating to Fig. 2 of the main text, we produce a 'heat map', Fig. 10, for  $\beta^{\text{eff}}$ , and for  $Q^*$  (defined in main text), for each machine, which shows the number of instances found in each small region. We notice that in the upper figure that the larger the effective inverse temperature, more instances fall within the 'thermal region', indicating a stronger dependence on the temperature for these instances. These instances correspond to the ones which freezeout at a later point in the anneal, and thus a smaller

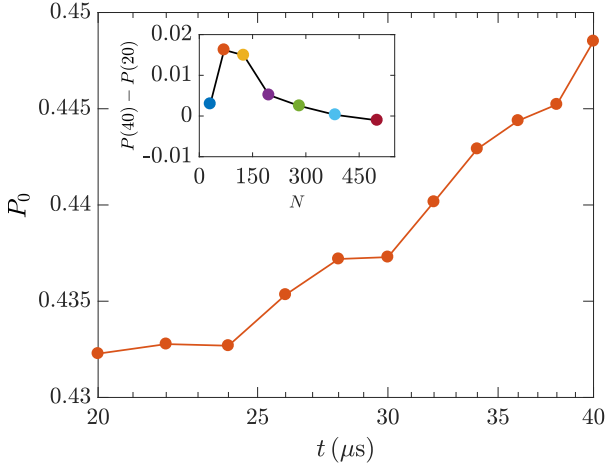


FIG. 8. **D-Wave success with anneal time, for single problem size.** Average probability of success,  $P_0$  (for the hotter USC machine), against anneal time (log scale), for problem size 70 ( $L = 3$ ). Note the approximate linear relationship. Inset: Difference in success probability between  $t = 40$  and  $t = 20 \mu s$  as measured by  $P(40) - P(20)$ , where  $P(t)$  is defined as  $P_0$  for anneal time  $t$ .

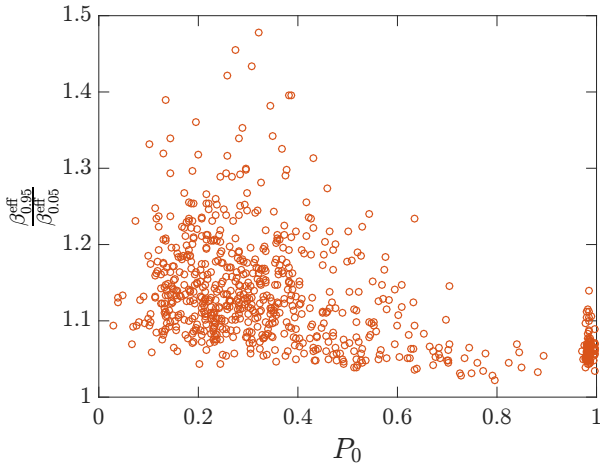


FIG. 9. **Variance with success probability.** Spread of effective inverse temperature as a function of success probability, as determined by the 95th to 5th percentile ratio of  $\beta^{\text{eff}}$  over all programming cycles, for  $N = 70$  (we pick this one as it is the problem size for which we have the most number of instances with reliable data).

$Q^*$  in the lower figure. In this lower figure, we observe the fit is closer to the 'ideal'  $y = x$  for smaller  $Q^*$ , and it deviates above this line for larger  $Q^*$  (we discuss in more detail in the main text).

Similarly relating to Fig. 2 of the main text, we produce Figs. 11, 12, which plot  $\beta^{\text{eff}}$ , and  $Q^*$ , for each machine, but split up by problem size. One can see that

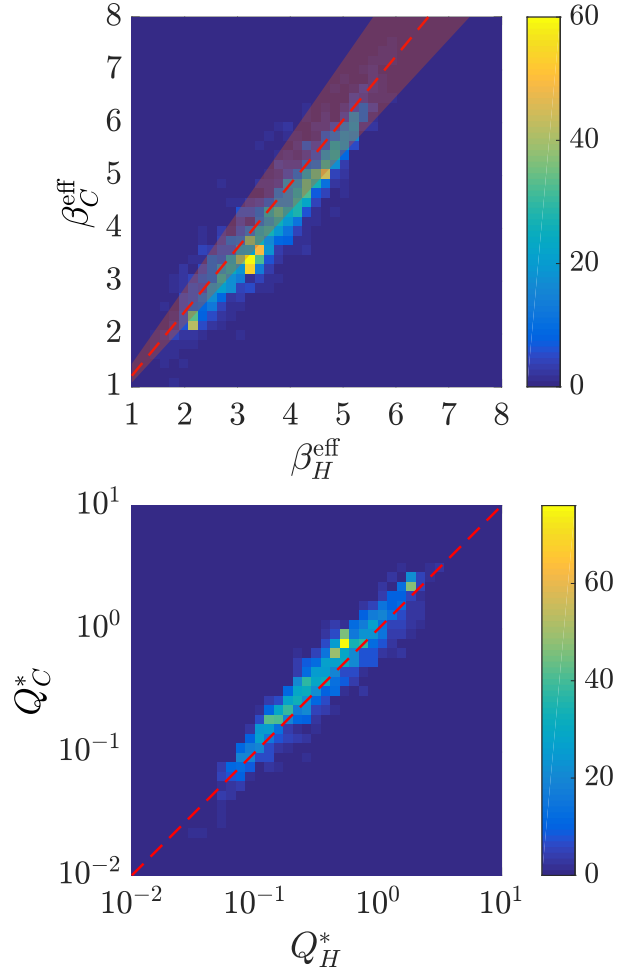


FIG. 10. **Density plots for  $\beta^{\text{eff}}$  and  $Q^*$ .** The corresponding 'heat map' for Fig. 2 of the main text. Color indicates the number of points in each region (given by color bar on the right hand side). Top: Red dash line is the 'ideal' thermal ratio (i.e. if the ratio of the effective inverse temperatures were the same as the physical 'thermal' inverse temperatures). The variance in physical temperature fluctuations is given by the red semi-transparent region. Bottom: Red dash line is  $y = x$ .

typically the larger problems exhibit lower values of  $\beta^{\text{eff}}$ , and likewise, larger values of  $Q^*$ , indicating these are in fact not thermalizing according to a Boltzmann distribution.

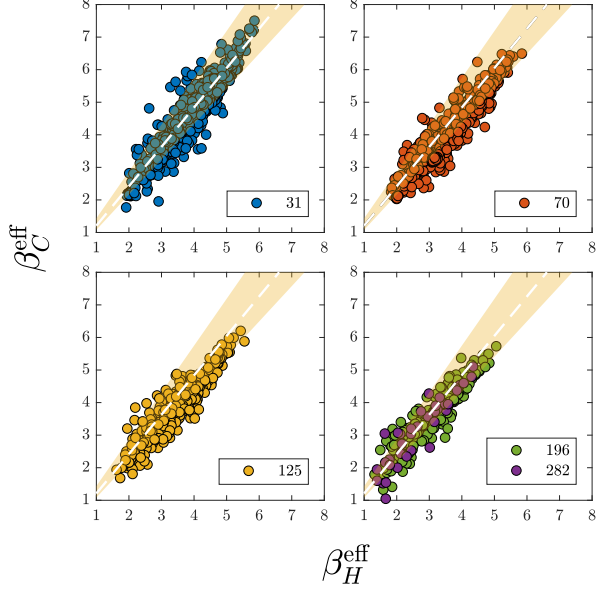


FIG. 11.  $\beta_C^{\text{eff}}$  for different problem sizes. These plots show the effective inverse temperature found by each machine (and for each instance), for the 5 different problem sizes (see legend) for which we have reliable degeneracy data. White dash line is the 'ideal' thermal ratio (i.e. if the ratio of the effective inverse temperatures were the same as the physical 'thermal' inverse temperatures). The variance in physical temperature fluctuations is given by the yellow semi-transparent region.

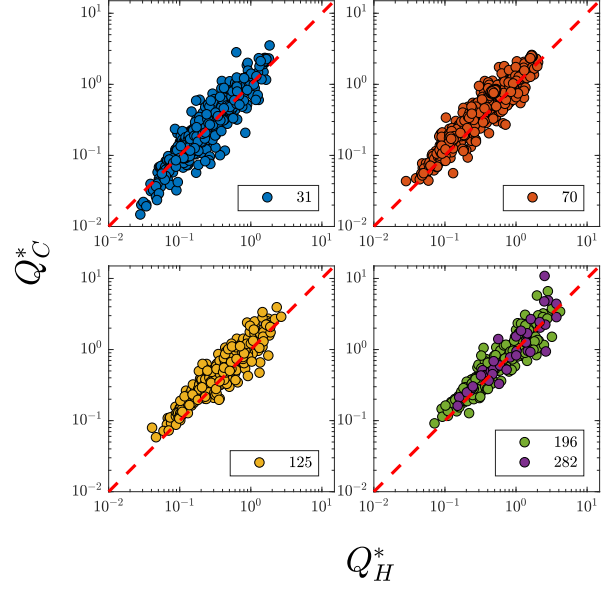


FIG. 12.  $Q^*$  for different problem sizes. These plots show the median  $Q^*$  found by each machine (and for each instance), for the 5 different problem sizes (see legend) for which we have reliable degeneracy data. Red dash line is  $y = x$ .